Beating the Correlation Breakdown, for Pearson's and Beyond: Robust Inference and Flexible Scenarios and Stress Testing for Financial Portfolios

Post 1 of 4: INTRODUCTION

JD Opdyke, Chief Analytics Officer, Partner, Sachs Capital Group Asset Management, LLC
JDOpdyke@gmail.com

NOTE: These posts summarize a chapter in my forthcoming monograph for Cambridge University Press.

Introduction

We live in a multivariate world, and effective modeling of financial portfolios, including their construction, allocation, forecasting, and risk analysis, simply is not possible without explicitly modeling the dependence structure of their assets.

Many different measures of dependence structure are widely used, including the foundational Pearson's product moment correlation matrix, rank-based measures like Kendall's Tau and Spearman's Rho, as well as measures designed to capture highly non-linear dependence such as the tail dependence matrix, Chatterjee's correlation, Lancaster's correlation, and Szekely's distance correlation and their many variants.

While dependence structure can drive portfolio results more than many other parameters in investment and risk models – sometimes even more than their combined effects –the literature provides relatively little to define the finite-sample distributions of these dependence measures under challenging, real-world data conditions. Yet this is exactly what is needed to make valid inferences about their estimates, and to use these inferences for a myriad of essential purposes, such as hypothesis testing, dynamic monitoring, realistic and granular scenario and reverse scenario analyses, and mitigating the effects of correlation breakdowns during market upheavals (which is when we need valid inferences the most).

This is the Introduction to a series of four posts that present a straightforward method– Nonparametric Angles-based Correlation ("NAbC") – for defining the finite-sample distributions of a very wide range of dependence measures for portfolio analysis. The next post starts with a fully analytic solution for a narrow but foundational case (with a link provided to an interactive, downloadable spreadsheet), and sequentially expands NAbC's application in each post to eventually cover ANY positive definite dependence measure (including and beyond those listed above). NAbC remains highly flexible and straightforward in its implementation, yet robustly unaffected and unrestricted by the distributional challenges of real-world financial returns (see 1. in pdf below).

Motivation for NAbC's development has been its effective application for real-world financial portfolios (as opposed to textbook settings), so the solution is characterized by seven critically necessary results that no other method provides simultaneously:

1. validity under challenging, real-world data conditions, with marginal asset distributions characterized by notably varying degrees of serial correlation, non-stationarity, heavy-tailedness, and asymmetry

2. application to ANY positive definite dependence measure, including, for example, Pearson's product moment correlation, rank-based measures like Kendall's tau and Spearman's rho, the kernel-based generalization of Szekely's distance correlation, and the tail dependence matrix, among others.

3. it remains "estimator agnostic," that is, valid regardless of the sample-based estimator used to estimate any of the above-mentioned dependence measures

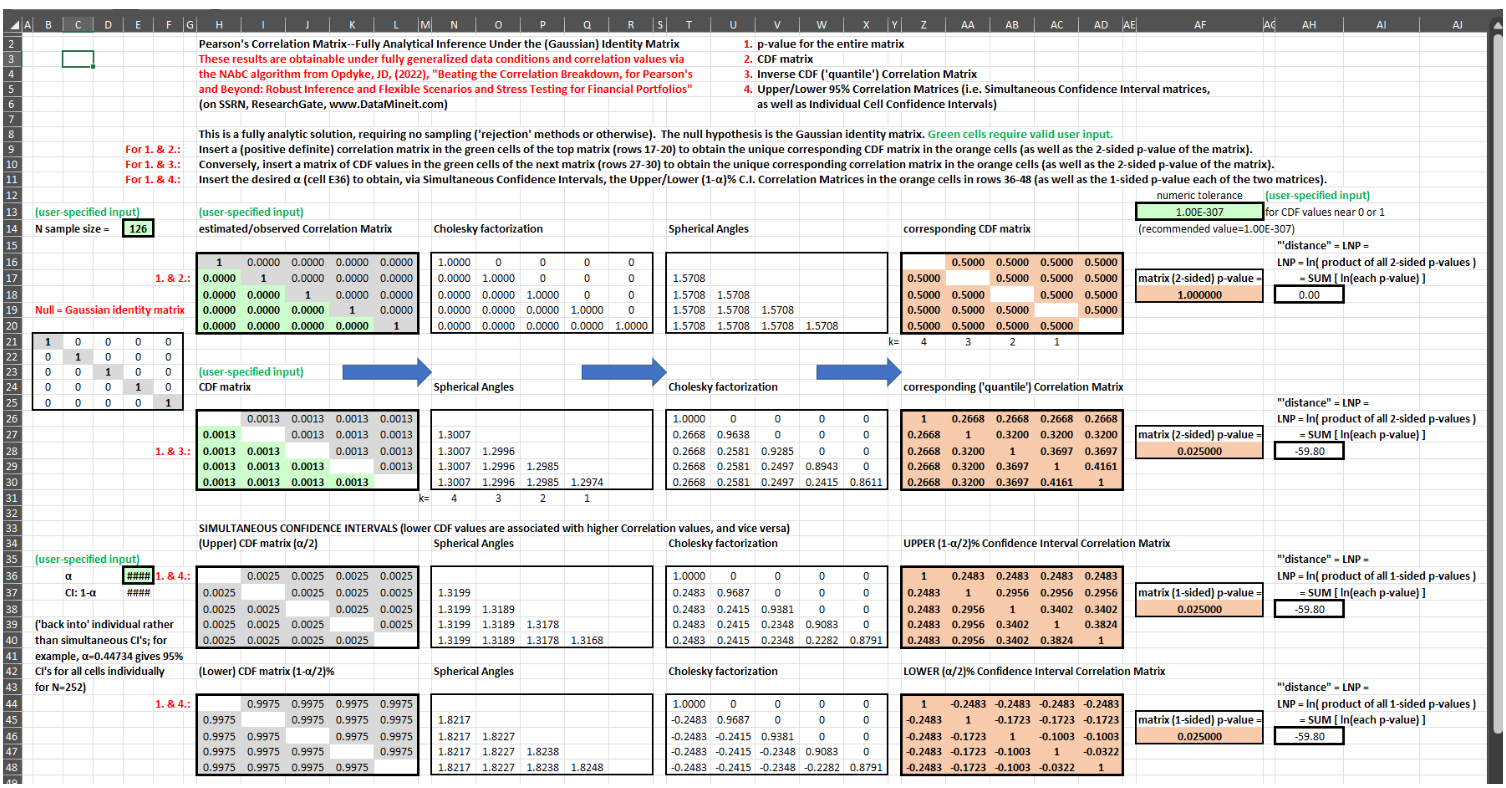
4. it provides valid confidence intervals and p-values at both the matrix-level and the pairwise cell-level, with analytic consistency between these two levels (ie the confidence intervals for all the cells define that of the entire matrix, and the same is true for the p-values; this effectively facilitates attribution analyses)

5. it provides a one-to-one quantile function, translating a matrix of all the cells' cdf values to a (unique) correlation (dependence measure) matrix, and back again, enabling precision in reverse scenarios and stress testing

6. all the above results remain valid even when selected cells in the matrix are 'frozen' for a given scenario or stress test, enabling granular and realistic scenarios

7. it remains valid not just asymptotically, ie for sample sizes presumed to be infinitely large, but rather, for the specific sample sizes we have in reality, enabling reliable application in actual, imperfect, non-textbook settings

To date, financial portfolio analysis in practice very often relies on ad hoc, largely qualitative, and 'judgmental' approaches to specifying and utilizing dependence structure, and when quantitative approaches are used, their valid application largely has been restricted to narrow cases. But practitioners, academics, and regulators have a long history of bringing analytic and probabilistic rigor to bear when estimating and analyzing the other parameters of our portfolio risk and investment models. Given that dependence structure often drives our portfolio results as much or more than many of those parameters, how can we settle for anything less than this same level of rigor when it comes to modeling our dependence structure? Tune in to the next post for an answer to this (semi)rhetorical question.



Pearson's Correlation Matrix--Fully Analytical Inference Under the (Gaussian) Identity Matrix
These results are obtainable under fully generalized data conditions and correlation values via the NAbC algorithm from Opdyke, JD, (2022), "Beating the Correlation Breakdown, for Pearson's and Beyond: Robust Inference and Flexible Scenarios and Stress Testing for Financial Portfolios" (on SSRN, ResearchGate, www.DataMineit.com)

1. p-value for the entire matrix
2. CDF matrix
3. Inverse CDF ('quantile') Correlation Matrix
4. Upper/Lower 95% Correlation Matrices (i.e. Simultaneous Confidence Interval matrices, as well as Individual Cell Confidence Intervals)

This is a fully analytic solution, requiring no sampling ('rejection' methods or otherwise). The null hypothesis is the Gaussian identity matrix. Green cells require valid user input.
For 1. & 2.: Insert a (positive definite) correlation matrix in the green cells of the top matrix (rows 17-20) to obtain the unique corresponding CDF matrix in the orange cells (as well as the 2-sided p-value of the matrix).
For 1. & 3.: Conversely, insert a matrix of CDF values in the green cells of the next matrix (rows 27-30) to obtain the unique corresponding correlation matrix in the orange cells (as well as the 2-sided p-value of the matrix).
For 1. & 4.: Insert the desired α (cell E36) to obtain, via Simultaneous Confidence Intervals, the Upper/Lower (1-α)% C.I. Correlation Matrices in the orange cells in rows 36-48 (as well as the 1-sided p-value each of the two matrices).

numeric tolerance (user-specified input): 1.00E-307 for CDF values near 0 or 1 (recommended value=1.00E-307)

(user-specified input) N sample size = 126

Null = Gaussian identity matrix

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

1. & 2.: (user-specified input) estimated/observed Correlation Matrix

1	0.0000	0.0000	0.0000	0.0000
0.0000	1	0.0000	0.0000	0.0000
0.0000	0.0000	1	0.0000	0.0000
0.0000	0.0000	0.0000	1	0.0000
0.0000	0.0000	0.0000	0.0000	1

Cholesky factorization

1.0000	0	0	0	0
0.0000	1.0000	0	0	0
0.0000	0.0000	1.0000	0	0
0.0000	0.0000	0.0000	1.0000	0
0.0000	0.0000	0.0000	0.0000	1.0000

Spherical Angles

1.5708			
1.5708	1.5708		
1.5708	1.5708	1.5708	
1.5708	1.5708	1.5708	1.5708

corresponding CDF matrix

	0.5000	0.5000	0.5000	0.5000
0.5000		0.5000	0.5000	0.5000
0.5000	0.5000		0.5000	0.5000
0.5000	0.5000	0.5000		0.5000
0.5000	0.5000	0.5000	0.5000	
k= 4	3	2	1	

matrix (2-sided) p-value = 1.000000

"'distance" = LNP = LNP = ln(product of all 2-sided p-values) = SUM [ln(each p-value)] = 0.00

1. & 3.: (user-specified input) CDF matrix

	0.0013	0.0013	0.0013	0.0013
0.0013		0.0013	0.0013	0.0013
0.0013	0.0013		0.0013	0.0013
0.0013	0.0013	0.0013		0.0013
0.0013	0.0013	0.0013	0.0013	

Spherical Angles

1.3007			
1.3007	1.2996		
1.3007	1.2996	1.2985	
1.3007	1.2996	1.2985	1.2974
k= 4	3	2	1

Cholesky factorization

1.0000	0	0	0	0
0.2668	0.9638	0	0	0
0.2668	0.2581	0.9285	0	0
0.2668	0.2581	0.2497	0.8943	0
0.2668	0.2581	0.2497	0.2415	0.8611

corresponding ('quantile') Correlation Matrix

1	0.2668	0.2668	0.2668	0.2668
0.2668	1	0.3200	0.3200	0.3200
0.2668	0.3200	1	0.3697	0.3697
0.2668	0.3200	0.3697	1	0.4161
0.2668	0.3200	0.3697	0.4161	1

matrix (2-sided) p-value = 0.025000

"'distance" = LNP = LNP = ln(product of all 2-sided p-values) = SUM [ln(each p-value)] = -59.80

SIMULTANEOUS CONFIDENCE INTERVALS (lower CDF values are associated with higher Correlation values, and vice versa)

(user-specified input)
α ####
CI: 1-α ####
('back into' individual rather than simultaneous CI's; for example, α=0.44734 gives 95% CI's for all cells individually for N=252)

1. & 4.: (Upper) CDF matrix (α/2)

	0.0025	0.0025	0.0025	0.0025
0.0025		0.0025	0.0025	0.0025
0.0025	0.0025		0.0025	0.0025
0.0025	0.0025	0.0025		0.0025
0.0025	0.0025	0.0025	0.0025	

Spherical Angles

1.3199			
1.3199	1.3189		
1.3199	1.3189	1.3178	
1.3199	1.3189	1.3178	1.3168

Cholesky factorization

1.0000	0	0	0	0
0.2483	0.9687	0	0	0
0.2483	0.2415	0.9381	0	0
0.2483	0.2415	0.2348	0.9083	0
0.2483	0.2415	0.2348	0.2282	0.8791

UPPER (1-α/2)% Confidence Interval Correlation Matrix

1	0.2483	0.2483	0.2483	0.2483
0.2483	1	0.2956	0.2956	0.2956
0.2483	0.2956	1	0.3402	0.3402
0.2483	0.2956	0.3402	1	0.3824
0.2483	0.2956	0.3402	0.3824	1

matrix (1-sided) p-value = 0.025000

"'distance" = LNP = LNP = ln(product of all 1-sided p-values) = SUM [ln(each p-value)] = -59.80

1. & 4.: (Lower) CDF matrix (1-α/2)%

	0.9975	0.9975	0.9975	0.9975
0.9975		0.9975	0.9975	0.9975
0.9975	0.9975		0.9975	0.9975
0.9975	0.9975	0.9975		0.9975
0.9975	0.9975	0.9975	0.9975	

Spherical Angles

1.8217			
1.8217	1.8227		
1.8217	1.8227	1.8238	
1.8217	1.8227	1.8238	1.8248

Cholesky factorization

1.0000	0	0	0	0
-0.2483	0.9687	0	0	0
-0.2483	-0.2415	0.9381	0	0
-0.2483	-0.2415	-0.2348	0.9083	0
-0.2483	-0.2415	-0.2348	-0.2282	0.8791

LOWER (α/2)% Confidence Interval Correlation Matrix

1	-0.2483	-0.2483	-0.2483	-0.2483
-0.2483	1	-0.1723	-0.1723	-0.1723
-0.2483	-0.1723	1	-0.1003	-0.1003
-0.2483	-0.1723	-0.1003	1	-0.0322
-0.2483	-0.1723	-0.1003	-0.0322	1

matrix (1-sided) p-value = 0.025000

"'distance" = LNP = LNP = ln(product of all 1-sided p-values) = SUM [ln(each p-value)] = -59.80